Global genetic variation of select opiate metabolism genes in self-reported healthy individuals

Wendt, F. R.

2018-03

Wendt , F R , Pathak , G , Sajantila , A , Chakraborty , R & Budowle , B 2018 , ' Global genetic variation of select opiate metabolism genes in self-reported healthy individuals ' , Pharmacogenomics Journal , vol. 18 , no. 2 , pp. 281-294 . https://doi.org/10.1038/tpj.2017.13

http://hdl.handle.net/10138/301996
https://doi.org/10.1038/tpj.2017.13

unspecified
publishedVersion

The Pharmacogenomics Journal (2018) 18, 281–294

www.nature.com/tpj

ORIGINAL ARTICLE

Global genetic variation of select opiate metabolism genes in self-reported healthy individuals

FR Wendt[1], G Pathak[1], A Sajantila[2], R Chakraborty[1] and B Budowle[1,3,4]

CYP2D6 is a key pharmacogene encoding an enzyme impacting poor, intermediate, extensive and ultrarapid phase I metabolism of many marketed drugs. The pharmacogenetics of opiate drug metabolism is particularly interesting due to the relatively high incidence of addiction and overdose. Recently, trans-acting opiate metabolism and analgesic response enzymes (*UGT2B7*, *ABCB1*, *OPRM1* and *COMT*) have been incorporated into pharmacogenetic studies to generate more comprehensive metabolic profiles of patients. With use of massively parallel sequencing, it is possible to identify additional polymorphisms that fine tune, or redefine, previous pharmacogenetic findings, which typically rely on targeted approaches. The 1000 Genomes Project data were analyzed to describe population genetic variation and statistics for these five genes in self-reported healthy individuals in five global super- and 26 sub-populations. Findings on the variation of these genes in various populations expand baseline understanding of pharmacogenetically relevant polymorphisms for future studies of affected cohorts.

The Pharmacogenomics Journal (2018) **18,** 281–294; doi:10.1038/tpj.2017.13; published online 11 April 2017

HIGHLIGHTS

- An *in silico* genetic analysis of five opiate metabolism genes (*CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1*, and *COMT*) was performed to identify SNPs, INDELs, and/or copy number variants in general populations.
- Allele frequencies, observed and expected heterozygosities, test results for Hardy Weinberg Equilibrium, and pairwise linkage disequilibria for polymorphisms in the introns, exons, 3' and 5' untranslated regions, and promoter regions of five genes are reported for 2 504 unrelated healthy individuals from five super-populations and 26 sub-populations.
- Multidimensional scaling plots show substantial inter-super-population separation while sub-populations show variable degrees of clustering within super-populations.
- *CYP2D6 ** alleles were used to determine activity scores for each sample, potentially identifying poor, intermediate, extensive, and ultrarapid metabolizer phenotypes in a cohort of self-reported healthy individuals.
- Principle component analyses of *CYP2D6* extensive metabolizers indicate intra-metabolizer phenotype variation.

INTRODUCTION

Cytochrome P450, family 2, subfamily D, polypeptide 6 (CYP2D6) is a clinically significant enzyme responsible for ~30% of phase I metabolism of ~25% of marketed drugs.[1,2] Of particular interest is the enzyme's role in the conversion of pain medications to active metabolites, namely morphine.[3–5] The highly polymorphic nature of *CYP2D6* results in various metabolizer phenotypes (MP; poor (PM), intermediate (IM), extensive (EM) and ultra-rapid (UM)),[6–8] typically inferred from the diplotype of *CYP2D6* star (*) alleles (a haplotype of one or more polymorphisms along the length of the gene),[9] that have been associated with lack of therapeutic response, idiosyncratic responses, or even death.[10–12]

Comprehensive pharmacogenetic studies have shown that single-nucleotide polymorphisms (SNPs) in other opiate metabolism and pain relief pathway genes also confer variable degrees of enzyme activity.[13–17] These additional genes of interest include uridine diphosphate glucuronosyltransferase, family 1, polypeptide B7 (*UGT2B7*), adenosine triphosphate-binding cassette, subfamily B, number 1 (*ABCB1*), opioid receptor mu 1 (*OPRM1*) and catechol-O-methyltransferase (*COMT*). *UGT2B7* encodes an enzyme that converts morphine to morphine-6-glucuronide; these two compounds are the primary cause of the analgesic effect of opiates. *ABCB1* encodes p-glycoprotein (or multidrug resistance protein 1), a membrane-associated transporter responsible for the efflux of morphine from various organs. *OPRM1* encodes the primary receptor for signal transduction of the analgesic response. Finally, *COMT* encodes a protein that interacts with the opioid receptor mechanism to modulate pain response through catecholamine breakdown. Polymorphisms within these genes can impact opiate metabolism by altering the performance of their protein products, leading to non-effective treatment or clinical complications following opiate medication administration.[14,15]

Previous pharmacogenetic studies have focused on identifying common causal polymorphisms using genome-wide association studies (targeted SNP arrays and targeted massively parallel sequencing) to determine the MP of ante- and post-mortem patients.[17–19] While valuable, these methods fail to assess polymorphisms comprehensively in a target sequence on the individual and population levels. In addition, they hinder discovery of novel polymorphisms that may provide greater insight into phenotypic variability and subsequent resequencing of target loci

[1]Institute for Molecular Medicine, University of North Texas Health Science Center, Fort Worth, TX, USA; [2]Department of Forensic Medicine, University of Helsinki, Helsinki, Finland; [3]Center for Human Identification, University of North Texas Health Science Center, Fort Worth, TX USA and [4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia. Correspondence: FR Wendt, Department of Molecular and Medical Genetics, Institute for Molecular Medicine University of North Texas Health Science Center 3500 Camp Bowie Bolevard, CBH-250 Fort Worth, 76107 TX, USA.
E-mail: Frank.Wendt@my.unthsc.edu
Received 21 September 2016; revised 16 February 2017; accepted 21 February 2017; published online 11 April 2017

may be required for confirmation of allele calls.[20] Massively parallel sequencing of the full gene region may reveal additional variants, with reliable depth of coverage, which refine the current working knowledge of *CYP2D6* * alleles, for example, those which introduce premature stop codons before the defining polymorphisms of a * allele.

Pharmacogenetic population studies often control for presence of disease phenotype while placing less emphasis on demography and population substructure as contributing factors to variable allele distribution which may confer different metabolic profiles in populations.[10,21,22] Consequently, false positive associations may arise regarding the relationship between genotype and MP.[23]

Herein, an *in silico* study of the complete gene sequences of *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1*, *COMT* and their respective promoter regions was performed to identify novel SNPs, insertion/deletion (INDEL) polymorphisms and copy number variants (CNVs), define baseline population genetic variation, and identify potential phenotypic variability in opiate metabolism and pain relief. A summary is provided of population statistics, variant effect predictions, and clustering of super- and sub-populations based on SNPs, INDELs and CNVs in five genes whose protein products are associated with opiate metabolism. Finally, the distribution of *CYP2D6* * alleles in five super-populations and 26 sub-populations is shown which provides additional information regarding variability within the population of EMs.[24] These findings serve as substantial population genetic data for healthy cohorts which may guide the pharmacogenetics community towards studies involving comprehensive genetic screening.

MATERIALS AND METHODS

Gene and promoter regions were identified using GeneCards Human Gene Database.[25] Genotype data were obtained from 2504 unrelated healthy individuals whose sequence data were downloaded from Phase 3 of the 1000 Genomes Project using the University of California Santa Cruz (UCSC) Table Browser[26,27] and the appropriate hg19 reference genome coordinates for *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1*, *COMT* and their respective promoter regions. The 1000 Genomes Project reports data with sequence depth of coverage $\geq 4\times$.

Population genetic summary statistics and statistical tests were performed for five super-populations (African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS)) and 26 sub-population (Supplementary Table 1). Allele frequencies, observed and expected heterozygosity calculations, and tests for departures from Hardy–Weinberg equilibrium (HWE) and pairwise linkage disequilibrium (LD, assuming HWE) were performed using Genetic Data Analysis Software.[28] Allele frequency 95% confidence intervals were estimated using the normal approximation to the binomial method. Tests for HWE departures and pairwise LD were performed for super- and sub-populations due to the potential for loci meeting HWE expectations or pairwise loci linkage equilibrium in sub-populations but deviating from these expectations when pooled into super-populations.[29] Due to the size of *ABCB1* and *OPRM1* and the number of polymorphisms within each gene, computation constraints with software memory were experienced while performing all tests for pairwise LD between these polymorphisms (~17 million and ~23 million pairwise comparisons for *ABCB1* and *OPRM1*, respectively). Consequently, tests for pairwise LD for *ABCB1* and *OPRM1* polymorphisms were performed between HWE-deviating loci and all other loci. Both tests are sensitive to low frequency alleles and focusing on this subset of loci for pairwise LD testing, under the assumption of HWE, could indicate if the polymorphisms are subject to some selective pressures and/or genotyping errors as a result of the relatively low coverage of 1000 Genomes Project data.[30] Here we use 'linkage disequilibrium block' to describe a cluster of polymorphisms with significant deviations from pairwise LD with all other polymorphisms for a gene. Ensembl Variant Predictor (Release 84, March 2016)[31] and Sort Intolerant From Tolerant (SIFT)[32–36] were used to determine SIFT, Polymorphism Phenotyping v2 (PolyPhen-2),[37,38] and Protein Variant Effect Analyzer (PROVEAN)[39–41] variant effect predictions and scores for all identified polymorphisms. Intronic positions within 1000 bases of an exon were further analyzed using Human Splicing Finder (HSF).[42] Multidimensional scaling (MDS) plots and principal component analysis plots were generated in RStudio.[43]

CYP2D6 * alleles were assigned according to the presence of causal polymorphisms associated with known phenotype[9] and were used to assign activity scores and MP to each individual.[44] Haplotypes producing no amino acid changes and lacking causal intronic polymorphisms were considered *1; haplotypes conferring the combination of R296C and S486T amino acid changes but lacking any other amino acid change and intronic causal polymorphisms were considered *2. Individuals possessing *CYP2D6* * alleles with undetermined effects on activity (*22, *28 and *43, for example), or haplotypes that could not be associated with a * allele, were removed from MP analyses.

RESULTS

CYP2D6

Allele frequencies for 418 polymorphic loci (402 SNPs, 15 INDELs and one CNV) in the *CYP2D6* region for five super-populations and 26 sub-populations are listed in Supplementary Table 2. The average observed heterozygosity for 26 sub-populations was 0.0341 ± 0.102 with a range of 0.0253 ± 0.0836 (CHS) to 0.0439 ± 0.114 (GWD; Table 1 and Supplementary Table 3). When pooled, the average super-population observed heterozygosity was 0.0384 ± 0.0980 for AFR, 0.0337 ± 0.102 for AMR, 0.0281 ± 0.0918 for EAS, 0.0359 ± 0.107 for EUR and 0.0339 ± 0.107 for SAS (Table 1 and Supplementary Table 3). After Bonferroni correction ($P < 0.000120$), one locus in GBR (rs35742686), one locus in EAS (rs374153932) and four loci in AFR (rs78854695, rs28371705, rs28371703 and rs376217512) significantly deviated from HWE, all of which are less than that due to chance alone (that is, ~21; Table 2 and Supplementary Table 4).

After Bonferroni correction, sub-populations exhibited an average of 470 ± 90 significant pairwise LDs with a range of 331 (ASW) to 721 (KHV) significant pairwise LDs and 3693 AFR, 799 AMR, 1048 EAS, 1031 EUR and 933 SAS significant pairwise LDs were observed ($P < 5.74 \times 10^{-7}$), all of which are less than that due to chance alone (~4358 pairwise comparisons; Table 2 and Supplementary Figure 1). LD heat-maps of five super-populations (Supplementary Figure 2) show a cluster of six to seven polymorphisms (rs29001678 (AMR, EUR, SAS only), rs1081000, rs28695233, rs75276289, rs76312385, rs74644586 and rs1080996), which appear to form an LD block. There were an average of 44 ± 14 significant pairwise LDs between these seven polymorphisms and others within the gene, with a range of 33 (AMR) to 71 (AFR) significant pairwise LDs. This group of polymorphisms is found within *CYP2D6* intron 1 (hg19 positions 42526524–42526573) and do not alter CYP2D6 function; however, rs1080995, rs74644586 and rs76312385 are part of the *CYP2D6**21A haplotype and may be observed in any *CYP2D6* * allele with an intron 1 gene conversion with *CYP2D7* (*CYP2D6**11, *14B, *21B, *63, *73, *84, *88, *98, *102, *103, *104 and *105).[9]

MDS plots (Figure 1) were created using *CYP2D6* polymorphism pairwise genetic distances between super-populations and within super-populations (between sub-populations). There was substantial separation of the AFR and EAS populations from the cluster of AMR, EUR and SAS populations while sub-population clustering is quite diverse within each super-population.

Variant effect prediction for 418 *CYP2D6* polymorphisms was performed using SIFT, PolyPhen-2 and PROVEAN (Table 3 and Supplementary Table 5).[32–41] Individual polymorphisms were assigned to one of five categories based on their SIFT, PolyPhen-2 and PROVEAN scores: tolerated with no discrepancies (predictions are concordant), discrepancies but most likely tolerated (predictions are discordant but favor tolerance), discrepancies but most likely damaging (predictions are discordant but favor intolerance), damaging with no discrepancies (predictions are concordant) and conflicting results (only two scores are reported and their predictions are discordant). Summaries of their frequencies and distribution across each gene are shown in Table 3 and Figure 2a, respectively. Due to the potential for multiple alternate alleles at the

Table 1. Average super-population and sub-population observed (H_o) and expected (H_e) heterozygosities across 418 *CYP2D6*, 613 *UGT2B7*, 5986 *ABCB1*, 6831 *OPRM1* and 1007 *COMT* polymorphisms.

Gene	*Super-population*	*Average He*	*Average Ho*	*Sub-population*	*Average He*	*Average Ho*
CYP2D6	AFR	0.0429 ± 0.110	0.0384 ± 0.0980	YRI	0.0417 ± 0.110	0.0365 ± 0.0956
				LWK	0.0435 ± 0.110	0.0386 ± 0.0984
				GWD	0.0433 ± 0.111	0.0440 ± 0.114
				MSL	0.0420 ± 0.109	0.0370 ± 0.0949
				ESN	0.0424 ± 0.111	0.0404 ± 0.107
				ASW	0.0417 ± 0.108	0.0360 ± 0.0956
				ACB	0.0429 ± 0.112	0.0346 ± 0.0895
	AMR	0.0372 ± 0.114	0.0337 ± 0.102	MXL	0.0340 ± 0.105	0.0296 ± 0.0892
				PUR	0.0405 ± 0.120	0.0413 ± 0.127
				CLM	0.0386 ± 0.115	0.0317 ± 0.0922
				PEL	0.0324 ± 0.108	0.0296 ± 0.0983
	EAS	0.0308 ± 0.102	0.0281 ± 0.0918	CHB	0.0310 ± 0.101	0.0310 ± 0.100
				JPT	0.0329 ± 0.109	0.0298 ± 0.0995
				CHS	0.0296 ± 0.0980	0.0253 ± 0.0836
				CDX	0.0288 ± 0.0955	0.0260 ± 0.0843
				KHV	0.0275 ± 0.0910	0.0282 ± 0.0955
	EUR	0.0400 ± 0.121	0.0359 ± 0.107	CEU	0.0410 ± 0.122	0.0353 ± 0.104
				TSI	0.04070 ± 0.123	0.0373 ± 0.112
				FIN	0.0376 ± 0.1160	0.0357 ± 0.111
				GBR	0.0402 ± 0.121	0.0320 ± 0.0949
				IBS	0.0401 ± 0.121	0.0386 ± 0.117
	SAS	0.0374 ± 0.118	0.0339 ± 0.107	GIH	0.0381 ± 0.121	0.0362 ± 0.115
				PJL	0.0340 ± 0.111	0.0333 ± 0.108
				BEB	0.0371 ± 0.1130	0.0312 ± 0.0949
				STU	0.0374 ± 0.119	0.0309 ± 0.0975
				ITU	0.0381 ± 0.121	0.0374 ± 0.119
UGT2B7	AFR	0.0573 ± 0.117	0.0582 ± 0.121	YRI	0.0530 ± 0.109	0.0554 ± 0.115
				LWK	0.0610 ± 0.125	0.0668 ± 0.140
				GWD	0.0524 ± 0.110	0.0503 ± 0.109
				MSL	0.0495 ± 0.103	0.0492 ± 0.105
				ESN	0.0604 ± 0.124	0.0663 ± 0.140
				ASW	0.0605 ± 0.125	0.0681 ± 0.143
				ACB	0.0639 ± 0.134	0.0551 ± 0.115
	AMR	0.0675 ± 0.150	0.0613 ± 0.136	MXL	0.0621 ± 0.140	0.0694 ± 0.158
				PUR	0.0723 ± 0.161	0.0684 ± 0.151
				CLM	0.0741 ± 0.166	0.0653 ± 0.146
				PEL	0.0448 ± 0.105	0.0420 ± 0.104
	EAS	0.0611 ± 0.142	0.0644 ± 0.151	CHB	0.0646 ± 0.150	0.0847 ± 0.200
				JPT	0.0636 ± 0.145	0.0654 ± 0.149
				CHS	0.0605 ± 0.141	0.0698 ± 0.165
				CDX	0.0595 ± 0.139	0.0468 ± 0.111
				KHV	0.0570 ± 0.133	0.0529 ± 0.127
	EUR	0.0741 ± 0.168	0.0777 ± 0.177	CEU	0.0738 ± 0.169	0.0836 ± 0.193
				TSI	0.0745 ± 0.167	0.0834 ± 0.189
				FIN	0.0744 ± 0.168	0.0665 ± 0.150
				GBR	0.0726 ± 0.167	0.0725 ± 0.168
				IBS	0.0746 ± 0.168	0.0814 ± 0.184
	SAS	0.0720 ± 0.164	0.0740 ± 0.170	GIH	0.0727 ± 0.167	0.0744 ± 0.172
				PJL	0.0738 ± 0.165	0.0730 ± 0.165
				BEB	0.0701 ± 0.159	0.0731 ± 0.167
				STU	0.0719 ± 0.165	0.0780 ± 0.181
				ITU	0.0713 ± 0.164	0.0713 ± 0.166
ABCB1	AFR	0.0295 ± 0.0872	0.0294 ± 0.0873	YRI	0.0288 ± 0.0884	0.0287 ± 0.0885
				LWK	0.0309 ± 0.0909	0.0300 ± 0.0880
				GWD	0.0283 ± 0.0860	0.0296 ± 0.0914
				MSL	0.0303 ± 0.0875	0.0295 ± 0.0855
				ESN	0.0302 ± 0.0895	0.0300 ± 0.0903
				ASW	0.0279 ± 0.0847	0.0277 ± 0.0853
				ACB	0.0294 ± 0.0877	0.0297 ± 0.0893
	AMR	0.0209 ± 0.0771	0.0209 ± 0.0781	MXL	0.0202 ± 0.0783	0.0194 ± 0.0775
				PUR	0.0209 ± 0.0763	0.0219 ± 0.0812
				CLM	0.0215 ± 0.0779	0.0212 ± 0.0767
				PEL	0.0199 ± 0.0780	0.0205 ± 0.0821
	EAS	0.0186 ± 0.0758	0.0184 ± 0.0751	CHB	0.0177 ± 0.0733	0.0171 ± 0.0711
				JPT	0.0193 ± 0.0775	0.0196 ± 0.0795
				CHS	0.0192 ± 0.0779	0.0191 ± 0.0762
				CDX	0.0177 ± 0.0747	0.0182 ± 0.0789
				KHV	0.0188 ± 0.0769	0.0178 ± 0.0735

Table 1. (Continued)

Gene	*Super-population*	*Average He*	*Average Ho*	*Sub-population*	*Average He*	*Average Ho*
	EUR	0.0189 ± 0.0759	0.0192 ± 0.0780	CEU	0.0185 ± 0.0757	0.0193 ± 0.0807
				TSI	0.0195 ± 0.0771	0.0186 ± 0.0738
				FIN	0.0184 ± 0.0753	0.0188 ± 0.0785
				GBR	0.0182 ± 0.0762	0.0191 ± 0.0801
				IBS	0.0193 ± 0.0778	0.0201 ± 0.0817
	SAS	0.0174 ± 0.0688	0.0173 ± 0.0678	GIH	0.0175 ± 0.0706	0.0169 ± 0.0666
				PJL	0.0185 ± 0.0724	0.0185 ± 0.0723
				BEB	0.0170 ± 0.0677	0.0175 ± 0.0695
				STU	0.0165 ± 0.0658	0.0159 ± 0.0631
				ITU	0.0175 ± 0.0707	0.0174 ± 0.0713
OPRM1	AFR	0.0405 ± 0.101	0.0407 ± 0.102	YRI	0.0408 ± 0.104	0.0413 ± 0.106
				LWK	0.0412 ± 0.104	0.04100 ± 0.102
				GWD	0.0392 ± 0.101	0.0399 ± 0.105
				MSL	0.0380 ± 0.0968	0.0384 ± 0.0983
				ESN	0.0430 ± 0.108	0.0425 ± 0.107
				ASW	0.0390 ± 0.100	0.0414 ± 0.109
				ACB	0.0396 ± 0.100	0.0404 ± 0.103
	AMR	0.0299 ± 0.0949	0.0291 ± 0.0923	MXL	0.0302 ± 0.0982	0.0327 ± 0.108
				PUR	0.0313 ± 0.0953	0.0307 ± 0.0945
				CLM	0.0304 ± 0.0954	0.0309 ± 0.0983
				PEL	0.0244 ± 0.0852	0.0225 ± 0.0778
	EAS	0.0225 ± 0.0822	0.0228 ± 0.0835	CHB	0.0232 ± 0.083	0.0235 ± 0.0844
				JPT	0.0206 ± 0.0810	0.0210 ± 0.0824
				CHS	0.0235 ± 0.0834	0.0241 ± 0.0858
				CDX	0.0223 ± 0.0835	0.0228 ± 0.0873
				KHV	0.0226 ± 0.0829	0.0226 ± 0.0830
	EUR	0.0299 ± 0.0962	0.0302 ± 0.0980	CEU	0.0304 ± 0.0984	0.0302 ± 0.0987
				TSI	0.0290 ± 0.0939	0.0293 ± 0.0977
				FIN	0.0299 ± 0.0967	0.0315 ± 0.103
				GBR	0.0297 ± 0.0960	0.0292 ± 0.0957
				IBS	0.0304 ± 0.0981	0.0309 ± 0.0994
	SAS	0.0259 ± 0.0881	0.0258 ± 0.0888	GIH	0.0266 ± 0.0897	0.0265 ± 0.0901
				PJL	0.0256 ± 0.0880	0.0264 ± 0.0924
				BEB	0.0250 ± 0.0860	0.0245 ± 0.0851
				STU	0.0263 ± 0.0897	0.0267 ± 0.0916
				ITU	0.0254 ± 0.0887	0.0248 ± 0.0883
COMT	AFR	0.0489 ± 0.118	0.049 ± 0.118	YRI	0.0479 ± 0.118	0.0467 ± 0.114
				LWK	0.0493 ± 0.118	0.0479 ± 0.114
				GWD	0.0498 ± 0.121	0.0520 ± 0.128
				MSL	0.0484 ± 0.117	0.0473 ± 0.114
				ESN	0.0474 ± 0.117	0.0514 ± 0.131
				ASW	0.0503 ± 0.120	0.0498 ± 0.120
				ACB	0.0493 ± 0.120	0.0481 ± 0.117
	AMR	0.0453 ± 0.123	0.0442 ± 0.121	MXL	0.0442 ± 0.121	0.0462 ± 0.128
				PUR	0.0466 ± 0.125	0.0445 ± 0.120
				CLM	0.0461 ± 0.124	0.0472 ± 0.127
				PEL	0.0372 ± 0.111	0.0392 ± 0.123
	EAS	0.0429 ± 0.124	0.0425 ± 0.122	CHB	0.0442 ± 0.125	0.0423 ± 0.120
				JPT	0.0442 ± 0.124	0.0466 ± 0.131
				CHS	0.0411 ± 0.123	0.0420 ± 0.126
				CDX	0.0423 ± 0.123	0.0392 ± 0.115
				KHV	0.0424 ± 0.124	0.0418 ± 0.123
	EUR	0.0435 ± 0.122	0.0443 ± 0.125	CEU	0.0435 ± 0.123	0.0458 ± 0.130
				TSI	0.0441 ± 0.125	0.0467 ± 0.133
				FIN	0.0414 ± 0.115	0.0401 ± 0.112
				GBR	0.0437 ± 0.124	0.0436 ± 0.124
				IBS	0.0428 ± 0.122	0.0451 ± 0.129
	SAS	0.0456 ± 0.123	0.0437 ± 0.118	GIH	0.0463 ± 0.125	0.0460 ± 0.124
				PJL	0.0455 ± 0.124	0.0446 ± 0.123
				BEB	0.0448 ± 0.123	0.0404 ± 0.111
				STU	0.0459 ± 0.124	0.0417 ± 0.112
				ITU	0.0444 ± 0.121	0.0452 ± 0.126

Abbreviations: AFR, African; AMR, Ad Mixed American; ACB, African Caribbean in Barbados; ASW, American of African Ancestry in Southwest USA; BEB, Bengali from Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah Residence with Northern and Western Ancestry; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Colombia; EAS, East Asian; ESN, Esan in Nigeria; EUR, European; FIN, Finnish in Finland; GBR, British in England and Scotland; GIH, Gujarati Indian from Houston, Texas; GWD, Gambian in Western Divisions in Gambia; IBS, Iberian Population in Spain; ITU, Indian Telugu from the United Kingdom; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican Ancestry from Los Angeles, USA; PEL, Peruvians from Lima, Peru; PJL, Punjabi from Lahore, Pakistan; PUR, Puerto Ricans from Puerto Rico; SAS, South Asian; STU, Sri Lankan Tamil from the United Kingdom; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria.

Table 2. Number of loci that deviated from HWE expectations and the number of pairwise loci comparisons that exhibited LD for *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* polymorphisms in five super-populations and 26 sub-populations. Bonferroni corrected HWE *P*-values were 0.000120, 8.16×10^{-5}, 8.35 X 10^{-6}, 7.32×10^{-6} and 4.96×10^{-5} for *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1* and *COMT*, respectively; Bonferroni corrected pairwise LD *P*-values were 5.34×10^{-7}, 2.67×10^{-7}, 5.50×10^{-8}, 2.24×10^{-8} and 9.87×10^{-8} for *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1* and *COMT*, respectively.

Gene	*Super-population*	*Significant HWE deviations*	*Significant LDs*	*Sub-population*	*Significant HWE deviations*	*Significant LDs*
CYP2D6	AFR	4	3693	YRI	0	516
				LWK	0	500
				GWD	0	449
				MSL	0	452
				ESN	0	422
				ASW	0	331
				ACB	0	634
	AMR	0	799	MXL	0	383
				PUR	0	560
				CLM	0	504
				PEL	0	380
	EAS	1	1048	CHB	0	438
				JPT	0	385
				CHS	0	455
				CDX	0	425
				KHV	0	721
	EUR	0	1031	CEU	0	595
				TSI	0	494
				FIN	0	387
				GBR	1	575
				IBS	0	402
	SAS	0	933	GIH	0	402
				PJL	0	443
				BEB	0	472
				STU	0	512
				ITU	0	393
UGT2B7	AFR	4	7728	YRI	2	4403
				LWK	0	3643
				GWD	2	4271
				MSL	1	4053
				ESN	2	4711
				ASW	0	2671
				ACB	0	3546
	AMR	3	7282	MXL	0	2917
				PUR	0	3526
				CLM	0	3731
				PEL	1	3160
	EAS	2	5308	CHB	36	24 147
				JPT	1	3965
				CHS	2	4500
				CDX	1	4174
				KHV	1	4313
	EUR	3	6295	CEU	1	4153
				TSI	0	3793
				FIN	0	4332
				GBR	0	3743
				IBS	1	4159
	SAS	3	6574	GIH	0	3405
				PJL	2	3968
				BEB	1	3542
				STU	1	3962
				ITU	3	4959
ABCB1	AFR	9	72 978	YRI	0	11 405
				LWK	0	4972
				GWD	1	12 227
				MSL	2	14 988
				ESN	1	12 071
				ASW	0	2947
				ACB	1	13 847
	AMR	2	31 011	MXL	0	7170
				PUR	1	9362
				CLM	1	11 249
				PEL	0	5597
	EAS	5	37 802	CHB	2	15 053
				JPT	0	5892
				CHS	2	15 271
				CDX	0	6908
				KHV	1	9580
	EUR	2	26 637	CEU	2	10 442
				TSI	0	9939
				FIN	0	3123
				GBR	1	8771
				IBS	1	9135

Table 2. (Continued)

Gene	*Super-population*	*Significant HWE deviations*	*Significant LDs*	*Sub-population*	*Significant HWE deviations*	*Significant LDs*
	SAS	3	25 566	GIH	1	8190
				PJL	1	9611
				BEB	1	8979
				STU	1	10 653
				ITU	1	9323
OPRM1	AFR	12	172 560	YRI	2	36 581
				LWK	1	27 603
				GWD	4	47 005
				MSL	2	33 978
				ESN	0	24 996
				ASW	0	11 928
				ACB	1	18 034
	AMR	5	92 744	MXL	2	30 805
				PUR	1	31 564
				CLM	2	36 436
				PEL	0	60 103
	EAS	5	62 824	CHB	2	33 915
				JPT	4	38 296
				CHS	2	32 577
				CDX	2	23 930
				KHV	5	42 291
	EUR	6	76 181	CEU	3	36 491
				TSI	2	32 190
				FIN	1	33 169
				GBR	4	37 849
				IBS	1	22 631
	SAS	5	77 803	GIH	1	30 707
				PJL	4	41 472
				BEB	2	23 612
				STU	4	44 452
				ITU	3	33 269
COMT	AFR	1	7362	YRI	0	1421
				LWK	0	1428
				GWD	0	1252
				MSL	0	1003
				ESN	2	2492
				ASW	0	772
				ACB	0	1132
	AMR	2	7004	MXL	0	1196
				PUR	0	2068
				CLM	2	1669
				PEL	0	4661
	EAS	2	6712	CHB	0	2396
				JPT	0	1940
				CHS	0	1777
				CDX	0	1890
				KHV	1	3079
	EUR	3	7835	CEU	1	2229
				TSI	0	1685
				FIN	2	2123
				GBR	0	2162
				IBS	0	2391
	SAS	2	7502	GIH	0	2202
				PJL	0	1870
				BEB	0	3969
				STU	3	5326
				ITU	0	1874

Abbreviations: ACB, African Caribbean in Barbados; AFR, African; AMR, Ad Mixed American; ASW, American of African Ancestry in Southwest USA; BEB, Bengali from Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah Residence with Northern and Western Ancestry; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Colombia; EAS, East Asian; ESN, Esan in Nigeria; EUR, European; FIN, Finnish in Finland; GBR, British in England and Scotland; GIH, Gujarati Indian from Houston, Texas; GWD, Gambian in Western Divisions in Gambia; HWE, Hardy–Weinberg Equilibrium; IBS, Iberian Population in Spain; ITU, Indian Telugu from the United Kingdom; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LD, linkage disequilibrium; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican Ancestry from Los Angeles, USA; PEL, Peruvians from Lima, Peru; PJL, Punjabi from Lahore, Pakistan; PUR, Puerto Ricans from Puerto Rico; SAS, South Asian; STU, Sri Lankan Tamil from the United Kingdom; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria.

54 damaging, or most likely damaging, polymorphisms (locus rs1135830, for example, can produce a non-synonymous amino acid change or a premature stop codon), 47 single-amino acid changes, 4 premature stop codons, 2 frame-shift mutations, 1 CNV, 1 in-frame insertion and 1 in-frame deletion mutations would arise. Fifty percent (80/160) of the intronic and/or splice-associated polymorphisms were scored by HSF (Figure 2a and Supplementary Table 5). Seven of these loci (rs5030656, rs192358451, rs377504871, rs78854695, rs267608282, rs28371702 and rs267608275) were predicted to alter, or most likely alter, splicing of the gene. The locus rs28371702 is considered part of the haplotype for 35 * alleles although it has not been reported as functionally relevant.[9] The remaining six polymorphisms have not been reported as part of a recognized * allele. Interestingly, the four intronic polymorphisms

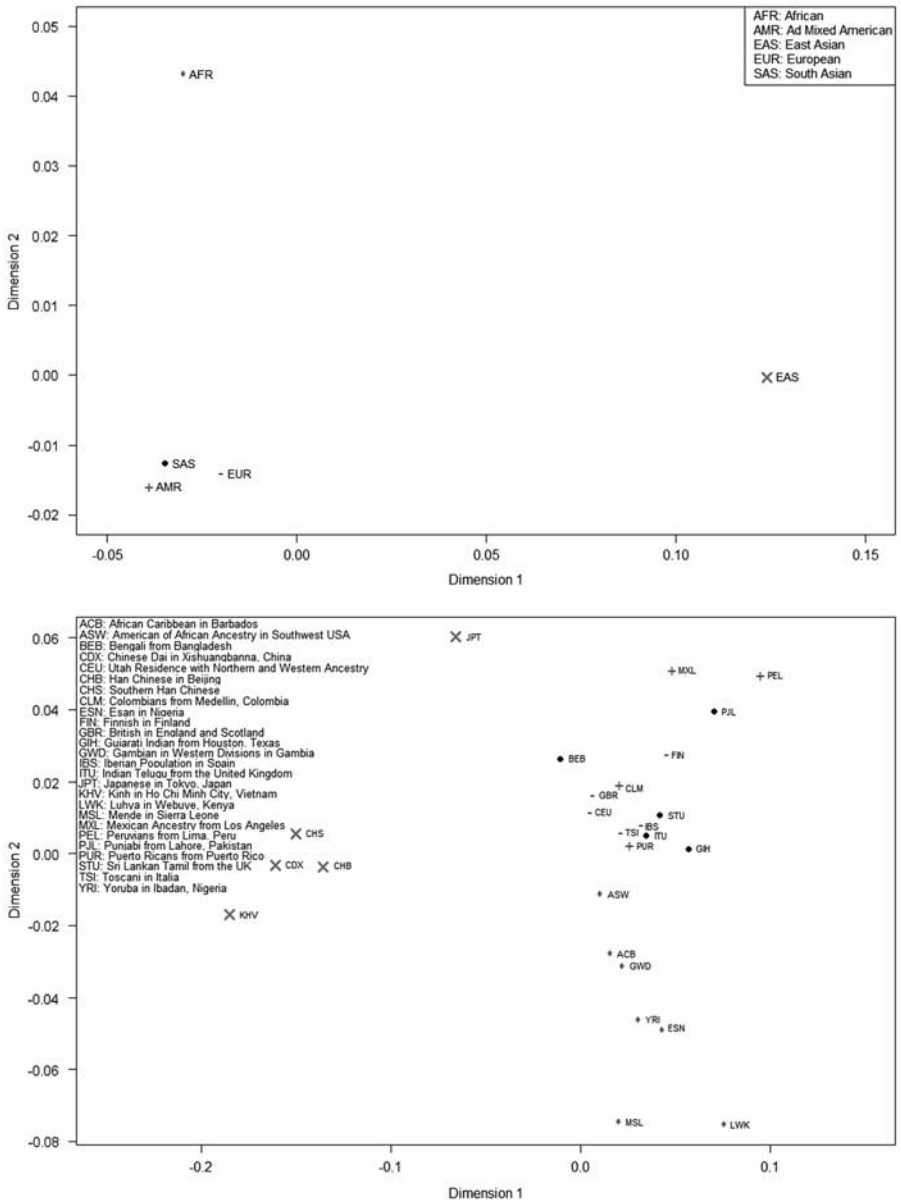


Figure 1. Multidimensional scaling plots of *CYP2D6* polymorphism pairwise genetic distances of five super-populations and 26 sub-populations based on 1000 Genome Project Phase 3 genotype data. African (AFR) populations are marked with a blue diamond, Ad Mixed American (AMR) populations are marked with a green plus sign, East Asian (EAS) populations are marked with a red 'X', European (EUR) populations are marked with a purple minus sign and South Asian (SAS) populations are marked with a solid black circle.

that are recognized by The Human Cytochrome p450 Allele Nomenclature Database[9] for causing splice-defects (883G>C [rs201377835], 1846G>A [rs3892097], 2950G>C (no rs number; invariable according to 1000 Genomes Project) and 2988G>A [rs28371725]) were either not scored by HSF or not considered variable sites in the 1000 Genomes Project and so genotypes were not exported from the UCSC Table Browser.

The Human CYP Allele Nomenclature Database[9] was used to assign * alleles to each sample. 210 unique haplotypes were observed in the 1000 Genomes Project Phase 3 data set, representing 37 * alleles (Supplementary Table 6). The average super-population observed and expected heterozygosities were 0.72 ± 0.080 and 0.78 ± 0.091, respectively. Using * allele assignments, *CYP2D6* significantly deviated from HWE expectations after Bonferroni correction in the AFR, AMR, EAS and SAS super-populations ($P < 0.0348$ for AFR and $P = 0.0420$, 0.0442 and 0.0348 in AMR, EAS and SAS, respectively) and seven sub-populations ($P = 0.000200$, 0.0277, 0.00290, 0.00510, 0.0202, 0.157 and 0.423 in ASW, LWK, MSL, YRI, CLM, British in England and Scotland and STU, respectively). After Bonferroni correction ($P = 0.01$ and $P = 0.0019$ for super- and sub-populations, respectively), the AFR super-population ($P < 0.01$) and ASW sub-population ($P = 0.000200$) significantly deviated from HWE expectations. Of the 210 observed haplotypes, only 14 (6.67%) are identical to those reported in the Human CYP Allele Nomenclature Table. Though not reported in the reference table, 84.8% of the remaining haplotypes could be associated with a * allele based on the presence of causal polymorphisms, however, 18 of them could not. These haplotypes represent 0.499% (25/5008) of the total 1000 Genomes Project haplotypes and contain

Table 3. Polymorphism effect categories for *CYP2D6*, *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* and promoter regions. Note that not all polymorphisms were assigned a score by each variant effect algorithm so the total counts for each algorithm may not equal the total of the other algorithms and may be different than the total number of polymorphisms for each gene (*N*).

Algorithm	*Effect category*	*CYP2D6* (N = *119*)		*UGT2B7* (N = *55*)		*ABCB1* (N = *94*)		*OPRM1* (N = *75*)		*COMT* (N = *45*)	
		Count	*Average score*	*Count*	*Average score*	*Count*	*Average score*	*Count*	*Average score*	*Count*	*Average score*
SIFT	Damaging	3	0.00900 ± 0.00870	0	–	0	–	10	0.000400 ± 0.00130	4	0.0165 ± 0.0158
	Deleterious	47	0.0157 ± 0.0147	17	0.0124 ± 0.0182	30	0.0160 ± 0.0167	6	0.00670 ± 0.103	3	0.00330 ± 0.00580
	Tolerated	63	0.634 ± 0.3707	33	0.666 ± 0.397	38	0.286 ± 0.239	46	0.324 ± 0.384	38	0.616 ± 0.364
PolyPhen-2	Probably damaging	16	0.978 ± 0.0241	5	0.963 ± 0.0322	5	0.9688 ± 0.0377	16	0.991 ± 0.0209	0	–
	Possibly damaging	17	0.743 ± 0.147	7	0.726 ± 0.0986	16	0.692 ± 0.117	5	0.682 ± 0.196	4	0.718 ± 0.194
	Benign	43	0.116 ± 0.129	22	0.0493 ± 0.0833	47	0.0505 ± 0.0714	21	0.0636 ± 0.0917	11	0.0939 ± 0.133
PROVEAN	Deleterious	52	− 4.89 ± 2.16	18	− 5.05 ± 2.41	30	− 4.90 ± 2.23	19	− 4.54 ± 1.56	5	− 5.20 ± 1.94
	Neutral	61	− 0.422 ± 0.978	37	− 0.204 ± 0.839	64	− 0.708 ± 0.851	56	− 0.0130 ± 0.518	40	− 0.186 ± 0.531

Polymorphism effect	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*
Damaging, no discrepancies	36	30.3	12	21.8	12	12.8	0	0	4	8.89
Discrepancies, most likely damaging	18	15.1	3	5.45	13	13.8	17	22.7	1	2.22
Discrepancies, most likely tolerated	10	8.40	5	9.09	19	20.2	13	17.3	1	2.22
Tolerated, no discrepancies	53	44.5	35	63.6	50	53.2	36	48.0	36	80.0
Conflicting results	2	1.68	0	0	0	0	9	12.0	3	6.67

Algorithm	*Effect category*	*CYP2D6* (N = *80*)		*UGT2B7* (N = *104*)		*ABCB1* (N = *564*)		*OPRM1* (N = *126*)		*COMT* (N = *84*)	
		Count	*Average score*	*Count*	*Average score*	*Count*	*Average score*	*Count*	*Average score*	*Count*	*Average score*
HSF	Alters	43	74.8 ± 6.15	62	74.2 ± 7.39	293	72.6 ± 9.14	64	70.7 ± 12.9	49	70.6 ± 14.0
			74.8 ± 6.16		74.4 ± 8.56		70.9 ± 9.08		70.3 ± 11.6		74.4 ± 9.51
			− 0.165 ± 6.13		1.50 ± 14.3		3.92 ± 26.6		6.16 ± 51.8		11.7 ± 36.3
			44.3 ± 10.4		47.7 ± 7.14		50.1 ± 15.5		52.3 ± 15.5		50.6 ± 12.1
	Creates	27	74.2 ± 6.88	22	73.3 ± 6.69	85	72.3 ± 7.80	40	70.8 ± 9.02	23	75.7 ± 7.89
			71.7 ± 79.8		55.6 ± 16.8		73.0 ± 118		49.9 ± 68.1		57.1 ± 42.3
	Breaks		73.5 ± 7.38		72.6 ± 9.44		72.3 ± 9.28		72.1 ± 10.1		74.9 ± 4.93
		29	43.8 ± 13.2	24	53.4 ± 13.1	151	51.8 ± 16.0	34	53.7 ± 16.8	16	48.6 ± 11.8
–			26.8 ± 15.7		24.4 ± 27.1		25.7 ± 30.7		23.2 ± 32.4		34.8 ± 16.2
	Activates cryptic site		35.2 ± 18.5		46.7 ± 0.445		51.6 ± 18.4		45.7 ± 6.29		44.2 ± 2.53
		3	75.2 ± 7.88	2	74.6 ± 1.05	126	72.8 ± 8.14	3	69.4 ± 3.15	3	71.0 ± 2.53
			182 ± 164		59.8 ± 3.77		79.58 ± 145.7		54.2 ± 22.2		60.85 ± 3.46

Polymorphism effect	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*	*Count*	*Frequency (%)*
Most likely effects splicing	4	5.00	2	1.92	127	22.5	3	2.38	3	3.57
Potentially effects splicing	3	3.75	9	8.65	171	30.3	13	10.3	8	9.52
Probably no effect on splicing	73	91.25	93	89.4	266	47.2	110	87.3	73	86.9

Abbreviations: HSF, Human Splicing Finder. SIFT, PolyPhen-2 and PROVEAN score cutoffs are 0.05, 0.5 and −2.5, respectively, for distinguishing between harmful and tolerated polymorphisms.[26–35] SIFT 'damaging' and 'deleterious' predictions, and PolyPhen-2 'probably damaging' and 'possibly damaging' predictions, are qualitative classifications indicating greater and lesser degrees of confidence, respectively, in the predicted damage caused by a polymorphism.[26–32] Average HSF scores are reference (hg19) consensus score, mutant consensus score and variation score.[42]

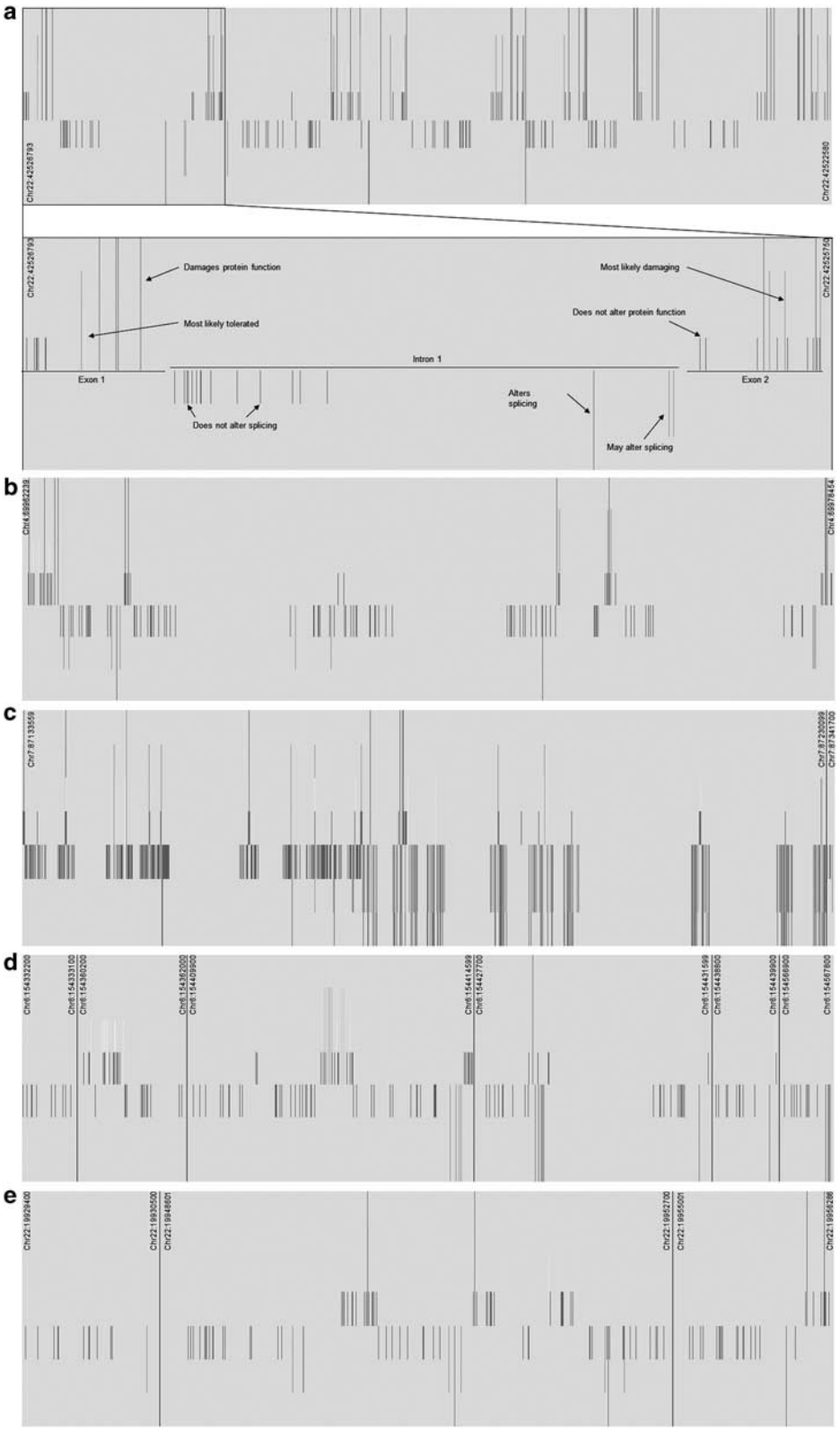


Figure 2. Qualitative summary of variant effect predictions. Each grey box represents a single gene: *CYP2D6* (**a**), *UGT2B7* (**b**), *ABCB1* (**c**), *OPRM1* (**d**) and *COMT* (**e**); the top vertical bars of each gene represent exonic polymorphisms scored by Sort Intolerant From Tolerant (SIFT), PolyPhen-2 and/or PROVEAN, the bottom bars represent intronic and splice-associated polymorphisms within 1000 bases of an exon that were scored by Human Splicing Finder (HSF), and black lines spanning both sections represent large unscored intronic regions that were removed; *CYP2D6* (**a**) and *UGT2B7* (**b**) are to scale while *ABCB1* (**c**), *OPRM1* (**d**) and *COMT* (**e**) have large intronic sequences (vertical black lines) removed; hg19 reference genome coordinates are provided.

Table 4. CYP2D6 metabolizer status counts and frequencies in 5 super-populations (bold) and 26 sub-populations based on available 1000 Genomes Phase 3 causative SNP genotype data. The number of individuals in each population is indicated in parentheses; 'Undetermined' metabolizer phenotype individuals contain at least one *CYP2D6** allele with unknown effect on enzyme activity.

Population	*Poor*		*Intermediate*		*Extensive*		*Ultrarapid*		*Undetermined*	
	Count	*Frequency*	*Count*	*Frequency*	*Count*	*Frequency*	*Count*	*Frequency*	*Count*	*Frequency*
AFR (661)	**9**	**0.0136**	**35**	**0.0530**	**564**	**0.853**	**0**	**0**	**53**	**0.0802**
ACB (96)	2	0.0208	6	0.0625	82	0.8542	0	0	6	0.0625
GWD (113)	1	0.00885	2	0.0177	103	0.912	0	0	7	0.0619
ESN (99)	1	0.0101	11	0.111	79	0.798	0	0	8	0.0808
MSL (85)	3	0.0353	2	0.0235	70	0.824	0	0	10	0.118
YRI (108)	0	0	5	0.0463	97	0.898	0	0	6	0.0556
LWK (99)	0	0	4	0.0404	84	0.848	0	0	11	0.111
ASW (61)	2	0.0328	5	0.0820	49	0.803	0	0	5	0.0820
AMR (347)	**10**	**0.0288**	**10**	**0.0288**	**291**	**0.839**	**0**	**0**	**36**	**0.104**
PUR (104)	6	0.0577	5	0.0481	81	0.779	0	0	12	0.115
CLM (94)	4	0.0426	4	0.0426	74	0.787	0	0	12	0.128
PEL (85)	0	0	0	0	78	0.918	0	0	7	0.0824
MXL (64)	0	0	1	0.0156	58	0.906	0	0	5	0.0781
EAS (504)	**0**	**0**	**13**	**0.0258**	**488**	**0.968**	**0**	**0**	**3**	**0.00595**
CHS (105)	0	0	3	0.0286	100	0.952	0	0	2	0.0190
CDX (93)	0	0	3	0.0323	89	0.957	0	0	1	0.0108
KHV (99)	0	0	5	0.0505	94	0.949	0	0	0	0
CHB (103)	0	0	2	0.0194	101	0.981	0	0	0	0
JPT (104)	0	0	0	0	104	1	0	0	0	0
EUR (503)	**29**	**0.0577**	**32**	**0.0636**	**433**	**0.861**	**0**	**0**	**9**	**0.0179**
CEU (99)	5	0.0505	9	0.0909	81	0.818	0	0	1	0.0101
GBR (91)	11	0.121	11	0.121	68	0.747	0	0	1	0.0110
IBS (107)	3	0.0280	2	0.0187	98	0.916	0	0	4	0.0374
TSI (107)	5	0.0467	7	0.0654	93	0.869	0	0	2	0.0187
FIN (99)	5	0.0505	3	0.0303	90	0.909	0	0	1	0.0101
SAS (489)	**10**	**0.0204**	**24**	**0.0491**	**441**	**0.902**	**2**	**0.00409**	**12**	**0.0245**
PJL (96)	1	0.0104	7	0.0729	87	0.906	0	0	1	0.0104
BEB (86)	2	0.0233	5	0.0581	76	0.884	0	0	3	0.0349
STU (102)	3	0.0294	4	0.0392	90	0.882	1	0.00980	4	0.0392
ITU (102)	3	0.0294	5	0.0490	90	0.882	1	0.00980	3	0.0294
GIH (103)	1	0.00971	3	0.0291	98	0.951	0	0	1	0.00971

Abbreviations: AFR, African; AMR, Ad Mixed American; ACB, African Caribbean in Barbados; ASW, American of African Ancestry in Southwest USA; BEB, Bengali from Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah Residence with Northern and Western Ancestry; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Colombia; EAS, East Asian; EUR, European; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GIH, Gujarati Indian from Houston, Texas; GWD, Gambian in Western Divisions in Gambia; IBS, Iberian Population in Spain; ITU, Indian Telugu from the United Kingdom; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican Ancestry from Los Angeles, USA; PEL, Peruvians from Lima, Peru; PJL, Punjabi from Lahore, Pakistan; PUR, Puerto Ricans from Puerto Rico; SAS, South Asian; STU, Sri Lankan Tamil from the United Kingdom; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria.

combinations of functionally relevant amino acid changes (Supplementary Table 6).

MP was assigned according to Gaedigk *et al.*[44] (Table 4). A χ^2 goodness-of-fit test indicated no significant differences between observed MP frequencies of 1000 Genomes Project super-population data and theoretical predictions ($P = 0.99$), previously reported values for general United States major population groups ($P = 0.54$),[45] and world populations (African, American, East Asian, European and South Central Asian; $P = 0.99$).[24]

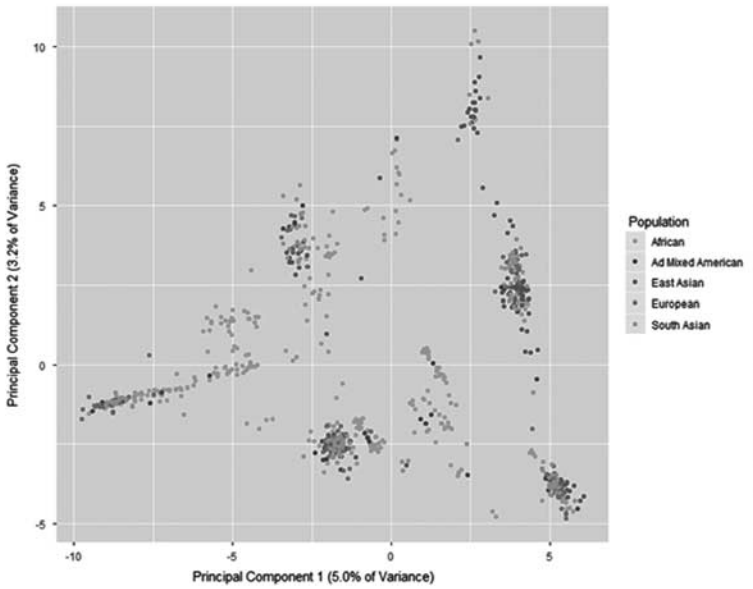
EM individuals were used to create principal component analysis plots by population (Figure 3). By super-population, the EM individuals display six prominent clusters with minimal overlap between AFR and EAS super-populations and considerable spread of the AMR, EUR and SAS populations across the entire plot. PC1 and PC2 explain greater than 5% of the variance for 10 and 8 polymorphisms, respectively. The same clustering pattern is observed for sub-populations with little clustering observed within populations (data not shown).

UGT2B7, *ABCB1*, *OPRM1* and *COMT*

Allele frequencies for 613 *UGT2B7* polymorphisms (585 SNPs and 28 INDELs), 5986 *ABCB1* polymorphisms (5775 SNPs, 210 INDELs and one CNV), 6831 *OPRM1* polymorphisms (6561 SNPs, 267 INDELs, 2 ALU element insertions and 1 CNV) and 1007 *COMT* polymorphisms (973 SNPs, 33 INDELs and one CNV) in 5 super-populations and 26 sub-populations are listed in Supplementary Tables 7–10.

The average super-population and sub-population observed and expected heterozygosities are listed in Table 1. A full list of each polymorphism and respective population-specific observed and expected heterozygosities are shown in Supplementary Tables 11–14.

A summary of the total number of polymorphisms in each gene and population that deviated from HWE expectations is listed in Table 2. A comprehensive list of HWE p-values for each polymorphism in each population is provided in Supplementary Tables 15–18. After Bonferroni correction, *UGT2B7* loci rs541550034 and rs57075995 ($P < 8.16 \times 10^{-5}$), *ABCB1* loci rs546527793 and rs570071012 ($P < 8.35 \times 10^{-6}$), and *OPRM1* loci rs147765820, rs376391508, rs77321666 and rs111829729 ($P < 7.32 \times 10^{-6}$) deviated from HWE expectations in all five super-populations. While no *COMT* loci deviated from HWE expectations in the five super-populations ($P = 4.97 \times 10^{-5}$), it should be noted that the loci rs138433986 and rs11912354 did deviate from HWE expectations



Locus	$Load_{PC1}$	$Load_{PC1}^2$	$Load_{PC2}$	$Load_{PC2}^2$	Functional Relevance?
rs28371730	0.226	0.051	0.006	0	-
rs1081000	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs28695233	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs75276289	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs74644586	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs1080996	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs1080995	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs76312385	0.225	0.051	0.002	0	Intron 1 conversion with CYP2D7
rs28624811	0.225	0.051	0.02	0	-
rs16947	0.225	0.05	0.019	0	2850C>T; C296R
rs4078247	0.091	0.008	0.272	0.074	-
rs28588594	0.093	0.009	0.272	0.074	-
rs1065852	0.092	0.008	0.271	0.074	100C>T; P34S
rs58440431	0.092	0.008	0.271	0.073	-
rs1080989	0.092	0.008	0.271	0.073	-
rs2004511	0.092	0.009	0.271	0.073	-
rs28371738	0.091	0.008	0.271	0.073	-
rs1081003	0.08	0.006	0.227	0.051	1039T>C; F112F

Figure 3. Principal component (PC) analysis of *CYP2D6* extensive metabolizers using genotypes of 418 polymorphisms from 1000 Genomes Project Phase 3. Samples are clustered according to super-population; rs numbers are provided for those loci best explained by PC1 and PC2; functional relevance of the polymorphism is indicated in reference to The Human Cytochrome p450 Allele Nomenclature Table[9] and concordance with variant effect prediction generated by SIFT, PolyPhen-2, PROVEAN and HSF with green and red cells indicating tolerance and damage, respectively.

in the AMR, EAS, EUR and SAS populations ($P = 0.0009$ and 0.0009). One sub-population, CHB, exhibited more deviations from HWE expectations than that due to chance alone (that is, ~20).

A summary of the total number of pairwise loci comparisons that demonstrated significant LDs are listed in Table 2 and the distribution of LD *P*-values is shown in Supplementary Figures 3–6. After Bonferroni correction, sub-populations exhibited an average of 4683 ± 4004, 9489 ± 3368, $33\,303 \pm 9716$ and 2154 ± 1071 significant LDs for *UGT2B7*, *ABCB1*, *OPRM1* and *COMT*, respectively. Pairwise LD heat-maps of *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* polymorphisms in five major super-populations (Supplementary Figures 7–10) show no substantial linkage blocks.

In contrast to *CYP2D6*, the individual MDS plots for *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* show substantial separation for all super-populations (Figure 4). Within super-populations, sub-populations cluster relatively well with minimal overlap between super-populations. Considering the entire data set of ~15 000 polymorphisms, MDS plots of super-populations follow the pattern observed with single-gene plots. However, sub-populations do not show any clustering within their respective super-populations.

Variant effect prediction was performed on 613 *UGT2B7*, 5986 *ABCB1*, 6831 *OPRM1* and 1007 *COMT* polymorphisms to generate SIFT, PolyPhen-2 and PROVEAN scores (Supplementary Tables 19–22).[32–41] A summary of the average score and frequency of each variant effect is displayed in Table 3. Of the damaging, or most likely, damaging, exonic polymorphisms in *UGT2B7*, *ABCB1*, *OPRM1* and *COMT*, 100% (15/15, 25/25, 17/17 and 5/5 polymorphisms in *UGT2B7*, *ABCB1*, *OPRM1* and *COMT*, respectively) are the result of single-amino acid changes. Intronic polymorphisms were analyzed further using HSF (Table 3). Those most likely to alter splicing of *UGT2B7*, *OPRM1* and *COMT* account for $< 5\%$ of the total number of polymorphisms scored by HSF. The intronic polymorphisms of *ABCB1* predicted to most likely, or potentially, alter splicing account for over 50% of the total (Table 3). These polymorphisms are distributed across introns 1 through 16, with very few splice-altering polymorphisms occurring after intron 16 (Figure 2c). In addition, one *COMT* polymorphism was recognized by the variant effect predictors as a frame-shift mutation (rs563298832) but was not assigned a score by the three algorithms used. Manual inspection of the locus in IGV shows the CATT deletion within intron 5 so assignment as a frame-shift mutation is incorrect. The HSF algorithm did not score this locus either. It is possible that this intronic polymorphism is damaging to the resulting protein, however, this assumption is not supported or refuted by the data presented.

Intergenic linkage disequilibria

A total of 1349 polymorphisms across all five target genes were assigned SIFT, PolyPhen-2, PROVEAN and/or HSF scores. Tests for pairwise LD were performed on this subset of loci to address potential linkage disequilibria between polymorphisms that may alter the activity of multiple proteins. After Bonferroni correction (5.50×10^{-8}), 9573 AFR, 1328 AMR, 2517 EAS, 3134 EUR and 2583 SAS significant pairwise LDs were observed between polymorphic loci of different genes ($P < 0.0004$, Supplementary Table 23). The number of significant pairwise LDs is less than that due to chance alone (that is, ~45 461), however, those that contain two causal polymorphisms may be clinically significant. After removal of significant pairwise LDs containing loci which deviate from HWE expectations, there were 539, 12, 124, 282 and 128 significant pairwise LDs in the AFR, AMR, EAS, EUR and SAS populations, respectively, between polymorphic loci in different genes that are predicted to be damaging, or most likely damaging to the resulting protein (Figure 5). Two polymorphisms are part of 82.2, 98.4, 46.8 and 85.9% of these significant pairwise LDs within AFR, EAS, EUR and SAS, respectively (rs5885589 and rs677830). Rs5885589 is an *ABCB1* intronic polymorphism which breaks an existing splice site and activates a cryptic splice site just upstream of exon 17. Rs677830 is found within exon 4 of *OPRM1* and confers glutamine411stop in transcript variant 1B5. https://www.ncbi.nlm.nih.gov/nuccore/NM_001145286.2. The AMR population does not have a substantial percentage of pairwise LDs associated with a single polymorphism.

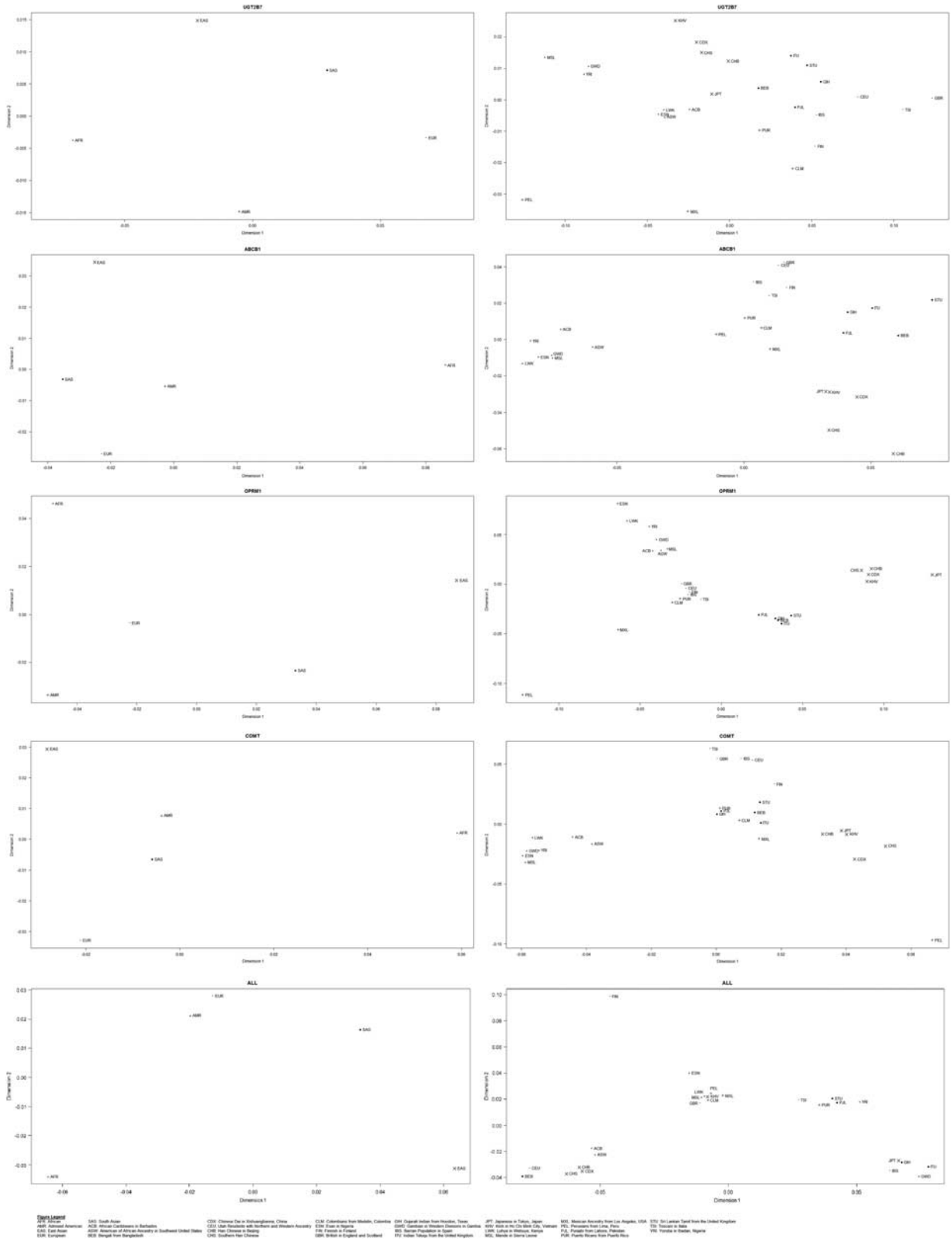


Figure 4. Multidimensional scaling plots of *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* polymorphism pairwise genetic distances of 5 super-populations and 26 sub-populations based on 1000 Genome Project Phase 3 genotype data. African (AFR) populations are marked with a blue diamond, Ad Mixed American (AMR) populations are marked with a green plus sign, East Asian (EAS) populations are marked with a red 'X', European (EUR) populations are marked with a purple minus sign and South Asian (SAS) populations are marked with a solid black circle.

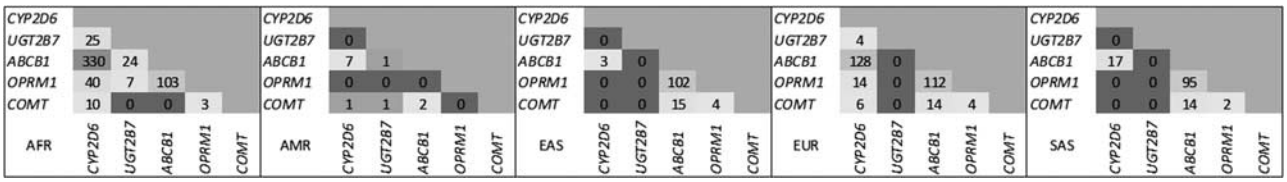


Figure 5. Summary of significant pairwise linkage disequilibria between polymorphisms on different genes in five major super-populations: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS).

DISCUSSION

Our study is limited by two factors. First, the coverage requirement for the 1000 Genomes Project is ~4×, producing an inherent level of missing variants or error in the sequence data. Second, due to limited size in each sub-population, some rare alleles may not be observed due to sample size. When data are generated in-house with greater sub-population samples sizes, greater coverage can be applied that will reduce the level of error and increase the chance of observing rare alleles. However, our analyses add to the population studies on pharmacogenetically interesting genes at global scale.[46–48]

Potential contributors to the number of significant deviations from HWE expectations that were observed for *CYP2D6* and *UGT2B7* polymorphisms in the ACB and CHB populations, respectively, are allele drop-out, the effects of selection and/or population substructure. For both sub-populations, some degree of substructure has been reported.[49–51] The Barbadian (ACB) population has demonstrated a higher degree of substructure relative to other ancestral African populations.[49,50] The Han Chinese also show some degree of substructure attributed to northern and southern Han populations. It has been shown that the 1000 Genomes CHB population contains individuals from these Han sub-groups.[51]

The 1000 Genomes Project contains self-reported healthy individuals and as such, the prevalence of *CYP2D6* PM, IM and UM metabolizers may not reflect previously published data sets focusing on cohorts of affected individuals. The principal component analysis plots of EMs explain relatively little variation (5.0 and 3.2%, respectively, for principle components one and two). These data support previous work demonstrating some level of intra-metabolizer status variability as well as intra-sub-population variability, which is supported by MDS plot of each population.

The *CYP2D6* MDS plots show separation of AFR and EAS from the cluster of AMR, EUR and SAS, supporting previously reported clinical differences between these populations.[52] Lack of tight sub-population (within super-population) clustering supports previous findings that *CYP2D6* activity variation may be greater within than between super-populations.[53] For example, the sub-populations within the EAS super-population (CDX, CHB, Southern Han Chinese, KHV and JPT) do not cluster tightly. The MDS plot indicates that the Chinese and Vietnamese populations (CDX, CHB, Southern Han Chinese and KHV) may be different from the Japanese (JPT) population. While minimal, this Asian variability is not novel and may be clinically significant when treating patients of these ancestries.[54] MDS plots of *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* show considerably less between super-population clustering, specifically of the SAS, EUR and AMR populations, suggesting that differences in these genes may be somewhat associated to super-populations. MDS plots of ~15 000 polymorphisms do not show sub-population clustering with their respective super-populations. This observation may be explained by the extreme allele frequency differences between sub-populations of the same super-population. For example, the *OPRM1* SNP, rs66579098, has alternate allele frequencies of 0.27, 0.33, 0.52 and 0.78 in the PUR, CLM, MXL and PEL sub-populations, respectively (belonging to the AMR super-population).[26,55]

Tests for pairwise LD of damaging, or likely damaging, polymorphisms in all five genes showed association between polymorphisms from all genes. The rs677830 (*OPRM1*) and rs5885589 (*ABCB1*) account for a substantial percentage of significant pairwise LDs in the AFR, EAS, EUR and SAS populations. These significant LDs may be clinically relevant due to the potential for multilocus interactions.[44] To our knowledge, rs677830 and rs5885589 have not been reported as causal polymorphisms. Interactions between these loci, or others, may be responsible for compensation when a damaging polymorphism dramatically alters normal protein activity, as suggested by Bartošová *et al*.[56] and Barratt *et al*.[57] with *ABCB1* and *OPRM1* polymorphisms shown to alter protein activity *in vivo*.

In conclusion, baseline population summary statistics are presented on five genes involved in opiate metabolism that have been implicated in phenotypic variability leading to idiosyncratic responses in patients. This study demonstrates some genetic association between *CYP2D6* and *UGT2B7*, *ABCB1*, *OPRM1* and *COMT* that will be important for future pharmacogenetic studies and combinatorial genetic approaches for patient care.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We would like to thank Emily Perry from the 1000 Genomes Helpdesk for assistance with extracting information from the Table Browser.

REFERENCES

1 Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol Ther* 2007; **116**: 496–526.
2 Ingelman-Sundberg M. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J* 2005; **5**: 6–13.
3 Leppert W. CYP2D6 in the metabolism of opioids for mild to moderate pain. *Pharmacology* 2011; **87**: 274-85.
4 Frost J, Helland A, Nordrum IS, Slørdal L. Investigation of morphine and morphine glucuronide levels and cytochrome P450 isoenzyme 2D6 genotype in codeine-related deaths. *Forensic Sci Int* 2012; **220**: 6–11.
5 Frost J, Løkken TN, Helland A, Nordrum IS, Slørdal L. Post-mortem levels and tissue distribution of codeine, codeine-6-glucuronide, norcodeine, morphine and morphine glucuronides in a series of codeine-related deaths. *Forensic Sci Int* 2016; **262**: 128–137.
6 Zhou SF, Di YM, Chan E, Du YM, Chow VD, Xue CC *et al.* Clinical pharmacogenetics and potential application in personalized medicine. *Curr Drug Metab* 2008; **9**: 738–784.
7 Sistonen J, Madadi P, Ross CJ, Yazdanpanah M, Lee JW, Landsmeer ML *et al.* Prediction of codeine toxicity in infants and their mothers using a novel combination of maternal genetic markers. *Clin Pharmacol Ther* 2012; **91**: 692–699.
8 Weber A, Szalai R, Sipeky C, Magyari L, Melegh M, Jaromi L *et al.* Increased prevalence of functional minor allele variants of drug metabolizing CYP2B6 and CYP2D6 genes in Roma population samples. *Pharmacol Rep* 2015; **67**: 460–464.

9 The Human Cytochrome p450 Allele Nomenclature Database. Accessed on May 2016. Available at http://www.cypalleles.ki.se/cyp2d6.htm.
10 Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, Belfer I *et al.* Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Hum Mol Genet* 2005; **14**: 135–143.
11 Koren G, Cairns J, Chitayat D, Gaedigk A, Leeder SJ. Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother. *Lancet* 2006; **368**: 704.
12 Sallee FR, DeVane CL, Ferrell RE. Fluoxetine-related death in a child with cytochrome P-450 2D6 genetic deficiency. *J Child Adolesc Psychopharmacol* 2000 Spring; **10**: 27–34.
13 Altar CA, Carhart JM, Allen JD, Hall-Flavin DK, Dechairo BM, Winner JG. Clinical validity: combinatorial pharmacogenomics predicts antidepressant responses and healthcare utilizations better than single gene phenotypes. *Pharmacogenomics J* 2015; **15**: 443–451.
14 Lam J, Woodall KL, Solbeck P, Ross CJ, Carleton BC, Hayden MR *et al.* Codeine-related deaths: The role of pharmacogenetics and drug interactions. *Forensic Sci Int* 2014; **239**: 50–56.
15 Baber M, Chaudhry S, Kelly L, Ross C, Carleton B, Berger H *et al.* The pharmacogenetics of codeine pain relief in the postpartum period. *Pharmacogenomics J* 2015; **15**: 430–435.
16 Bastami S, Gupta A, Zackrisson AL, Ahlner J, Osman A, Uppugunduri S. Influence of UGT2B7, OPRM1 and ABCB1 gene polymorphisms on postoperative morphine consumption. *Basic Clin Pharmacol Toxicol* 2014; **115**: 423–431.
17 Yuferov V, Levran O, Proudnikov D, Nielsen DA, Kreek MJ. Search for genetic markers and functional variants involved in the development of opiate and cocaine addiction and treatment. *Ann N Y Acad Sci* 2010; **1187**: 184–207.
18 Brion M, Sobrino B, Martinez M, Blanco-Verea A, Carracedo A. Massive parallel sequencing applied to the molecular autopsy in sudden cardiac death in the young. *Forensic Sci Int Genet* 2015; **18**: 160–170.
19 Narula N, Tester DJ, Paulmichl A, Maleszewski JJ, Ackerman MJ. Post-mortem Whole exome sequencing with gene-specific analysis for autopsy-negative sudden unexplained death in the young: a case series. *Pediatr Cardiol* 2015; **36**: 768–778.
20 Koch WH. Technology platforms for pharmacogenomic diagnostic assays. *Nat Rev Drug Discov* 2004; **3**: 749–761.
21 Brandl EJ, Tiwari AK, Zhou X, Deluce J, Kennedy JL, Müller DJ *et al.* Influence of CYP2D6 and CYP2C19 gene variants on antidepressant response in obsessive-compulsive disorder. *Pharmacogenomics J* 2014; **14**: 176–181.
22 Levo A, Koski A, Ojanperä I, Vuori E, Sajantila A, Post-mortem SNP. analysis of CYP2D6 gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood. *Forensic Sci Int* 2003; **135**: 9–15.
23 Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010; **11**: 356–366.
24 Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Klein T, Leeder JS. Prediction of CYP2D6 phenotype from genotype across world populations. *Genet Med* 2016; **19**: 69–76.
25 Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N *et al. In silico* human genomics with GeneCards. *Hum Genomics* 2011; **5**: 709–717.
26 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
27 Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. *Curr Protoc Bioinformatics* 2012; **Chapter 1**: Unit 1.4.
28 Weir BS. *1 Genetic Data Analysis*. 2nd edn. Sinauer Associates: Sunderland, MA, 1996, pp 376.
29 Wang J, Shete S. Testing departure from Hardy-Weinberg proportions. *Methods Mol Biol* 2012; **850**: 77–102.
30 Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M. On the usage of HWE for identifying genotyping errors. *Ann Hum Genet* 2007; **71**: 701–703.
31 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; **26**: 2069–2070.
32 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073–1081.
33 Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 2006; **7**: 61–80.
34 Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31**: 3812–3814.
35 Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 2002; **12**: 436–446.
36 Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001; **11**: 863–874.
37 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
38 Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013; **Chapter 7**: Unit 7.20.
39 Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 2012; **7**: e46688.
40 Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015; **31**: 2745–2747.
41 Choi Y (2012). A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-Locus Variants of Another Protein. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB '12). ACM, New York, NY, USA, 414–417.
42 Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009; **37**: e67.
43 RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, USA. doi: http://www.rstudio.com/.
44 Gaedigk A, Simon SD, Pearce RE, Bradford LD, Kennedy MJ, Leeder JS. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin Pharmacol Ther* 2008; **83**: 234–242.
45 Bernard S, Neville KA, Nguyen AT, Flockhart DA. Interethnic differences in genetic polymorphisms of CYP2D6 in the U.S. population: clinical implications. *Oncologist* 2006; **11**: 126–135.
46 Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG *et al.* Population genetic structure of variable drug response. *Nat Genet* 2001; **29**: 265–269.
47 Li J, Zhang L, Zhou H, Stoneking M, Tang K. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Hum Mol Genet* 2011; **20**: 528–540.
48 Mizzi C, Dalabira E, Kumuthini J, Dzimiri N, Balogh I, Başak N *et al.* A European Spectrum of Pharmacogenomic Biomarkers: Implications for Clinical Pharmacogenomics. *PLoS ONE* 2016; **11**: e0162866.
49 Murray T, Beaty TH, Mathias RA, Rafaels N, Grant AV, Faruque MU *et al.* African and non-African admixture components in African Americans and an African Caribbean population. *Genet Epidemiol* 2010; **34**: 561–568.
50 Benn-Torres J, Bonilla C, Robbins CM, Waterman L, Moses TY, Hernandez W *et al.* Admixture and population stratification in African Caribbean populations. *Ann Hum Genet* 2008; **72**: 90–98.
51 Xu S, Yin X, Li S, Jin W, Lou H, Yang L *et al.* Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet* 2009; **85**: 762–774.
52 Yasuda SU, Zhang L, Huang SM. The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. *Clin Pharmacol Ther* 2008; **84**: 417–423.
53 Sistonen J, Sajantila A, Lao O, Corander J, Barbujani G, Fuselli S. CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet Genomics* 2007; **17**: 93–101.
54 Qin S, Shen L, Zhang A, Xie J, Shen W, Chen L *et al.* Systematic polymorphism analysis of the CYP2D6 gene in four different geographical Han populations in mainland China. *Genomics* 2008; **92**: 152–158.
55 Sulovari A, Chen YH, Hudziak JJ, Li D. Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. *Hum Genet* 2017; **136**: 39–54.
56 Bartošová O, Polanecký O, Perlík F, Adámek S, Slanař O. OPRM1 and ABCB1 polymorphisms and their effect on postoperative pain relief with piritramide. *Physiol Res* 2015; **64**: S521–S527.
57 Barratt DT, Coller JK, Hallinan R, Byrne A, White JM, Foster DJ, Somogyi AA. ABCB1 haplotype and OPRM1 118 A > G genotype interaction in methadone maintenance treatment pharmacogenetics. *Pharmgenomics Pers Med* 2012; **5**: 53–62.

Supplementary Information accompanies the paper on the The Pharmacogenomics Journal website (http://www.nature.com/tpj)